Visualizations of SARS-CoV-2 Genomes on Genomic Index Maps

Jeffrey Zheng (conjugatelogic@yahoo.com)
Key Laboratory of Quantum Information of Yunnan, Key Laboratory of Software Engineering of Yunnan, Yunnan University https://orcid.org/0000-0003-4225-7077

Minghan Zhu
Yunnan University

Mu Qiao
Yunnan University

Yang Zhou
Yunnan University

Research Article

Keywords: genomic index, combinatorial entropy, integrated entropy, mean entropy, topological entropy, visualization, clustering, feature analysis

Posted Date: August 26th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-65159/v1

Abstract

In this paper, comprehensive cases are visualized by {C1, C2, C3, C4} four modules of the MAS. Four sets of SARS-CoV-2 genomes with 20, 128, 381 and 1337 cases were processed to be represented as a series of genomic index maps in four entropies: combinatorial, integrated, mean and topological entropies, respectively. Typical samples and explanations are described. Multiple levels of hierarchical representations are illustrated.

Introduction

Genomic Index for SARS-CoV-2 Genomes

The genomic index provides unique identification for each genome to be invari- ant under given conditions. Based on these types of global quantitative charac- teristics, it is convenient for huge numbers of genomes to be located in a certain geometric region to be collected as clusters. Four papers in this special issue dis- cuss this type of entropy quantity in separate papers: “Visualizations of Topological Entropy on SARS-CoV-2 Genomes in Multiple Regions”, “2D Graphical Analy- sis and Visualization of SARS-CoV-2”, “Cluster Analysis of Visual Differences on Pairs of SARS-CoV-2 Genomes”, “Visualizations of Combinatorial Entropy Index on Whole SARS-CoV-2 Genomes”.

Considering this is an extremely important research direction, it is necessary to handle this topic from a foundation level to provide additional information to ex- plore hidden structures among this type of multiple levels of hierarchical construc- tions from an integrated viewpoint [3]-[9].

Useful Visualizations

In the advanced fighting fields of COVID-19, various visualization schemes and in- teresting structures are widely illustrated in research papers and public websites. i.e. The well-known John Hopkins website for COVID-19, the newest genomes from GISAID viral databases and dynamic Phylogeny tree on Nextstrain attracted atten- tion from people worldwide for the newest progress and development on COVID-19 spreading situations. In relation to medical practices, intrinsic information of func- tional genomes provides refined differences for medical doctors and general prac- ticers to handle complicated viral toxicities, toxicology, pharmacology and so on. A research routine on viral toxicology is shown in Fig. 1. More than 270-fold tox- icities were identified among different SARS-CoV-2 genomes collected in various regions.

Brief Relationship between Phylogeny Trees and Genomic Index Maps

From a quantitative similarity viewpoint, the relationship is contained in phyloge- netic structures corresponding to density distributions represented in genomic index maps shown in Fig. 2. Based on this type of correspondence, a list of transforma- tions will be performed to show more complicated similarity information among various genomic index maps in the following parts.

Materials And Methods

Main Schemes in Processes

Following the architecture of the MAS, fifteen modules, input-output types and workflows are discussed in both papers: “A Visual Framework of Meta Genomic Analysis on Variations of Whole SARS-CoV-2 Sequences” and “Input-Output Types of Fifteen Modules on Discrete and Real Measurements for COVID-19”. Data flows, multiple and conditional probability measures, and key entropy equations are listed.

Specific processes are briefly discussed in the four separate papers on this issue.

All three projections: the *{*A, B, C*}* groups of the MAS are involved.

Datasets

All datasets used in this special issue from various open source genomic banks such as CNGBdb: Virus Database & SARS-CoV-2 Database; Database Commons; NCBI GenBank SARS-CoV-2-seqs and GISAID + GitHub + Nextstrain et al. More than two thousand whole sequences of SARS-CoV-2 over 200 countries have been col- lected mainly from the NCBI and GISAID site under the Nextstrain project from January to March 2020.

In addition, a set of coronaviruses, H1N1 virus, Bats & Pangolins, MERS, Ebola, SARS and other sequences were collected as samples for comparisons. One simi- larity comparison is shown in Fig. 3.

Three special datasets are collected to be contained in 128, 381 and 1337 samples of SARS-CoV-2 from multiple regions worldwide to illustrate specific effects under different combinatorial and other complicated conditions.

Results And Discussion

Visual Results

Sample Results

For 381 samples, the C1 genomic index maps are shown in Fig. 6 to Fig. 16 to illustrate complicated regions and various countries under different selected condi- tions, distinct combinations and projections.A list of sample results on various modules {C1, C2, C3, C4} are selected for illustrations as follows. Four datasets of SARS-CoV-2 genomes are involved: 20 (G20), 128, 381 and 1337. Genomic index maps for (C3,C2) are shown in Figs. 4 and 5 to illustrate variations in genomic index maps from the G20 to 128 genomes, respectively.

For 1331 genomes, the C4 genomic index maps are illustrated in Fig. 17 to Fig. 20. Five and eight countries are selected for comparisons under topological entropies.

It is convenient to check each set of genomic index maps to observe specific distributions and typical structures for COVID-19 patients worldwide.

Discussion

Three sets of genomic index maps consistently provide huge invaluable informa- tion to be extracted from different sizes of SARS-CoV-2 genomes worldwide under unique quantitative invariants.

G20 and 128 Genomes on C2 and C3 Genomic Index

In Fig. 4, the first set of 20 samples is selected from the G20 regions to be repre- sented in the (C3,C2) genomic index maps. The leftside map is shown in the G20 samples under k-mer conditions with multiple uncertainties for each sample, and the leftside map is shown in a much clearer map with distinct positions for each region by integrated mean operations. Gradually adding more genomes increased to 62 and 128, leftside maps were shown in similar clustering patterns, and 128 maps showed stronger clusters there. The integrated map is shown in more sample points, espe- cially clustered around left-bottom parts in the map. Basic distributions between 62 and 128 were similar in shape. Two integrated maps are shown in Fig. 5, and there are significant differences between G20 and 128.

381 Genomes on C1 Genomic Index

Fig. 6 to Fig. 16 provide a large number of different combinations and projections for the C1 genomic index maps. Under 2D-(A,G) projection, six 2D maps are shown in Fig. 6 to represent two integrated maps for four countries and refined regions.

In Fig. 7, both 1D-A and 1D-P projections are illustrated, leftsides on 1D-A and rightsides on 1D-G. Significant differences can be identified.

In Fig. 8, one 2D-(A,G) map and two 1D-A/1D-G projections are illustrated. There are clustering numbers associated with the relevant entropy value correspond- ing to whole distributions under conditions. In Fig. 9, the same maps as in Fig. 8 are arranged by their X and Y projections for convenient visual comparison, with the vertical distribution Y organized in horizontal directions as the same as the distribu- tions in the X directions.

In Fig. 10, both 2D-(A,G) and 1D-A maps provided for four counties, and each country was associated with a 1D-A map. In Fig. 11, both 2D-(A,G) and 1D-G maps provided for four counties, and each country was associated with a 1D-G map.

In Fig 12, one 2D-(A,G) USA, three regions in the USA(CA, NY and WA) and three regions in China (WH, SH and HZ) were selected to show each 1D-A map. In Fig 13, one 2D-(A,G) USA, three regions in the USA (CA, NY and WA) and three regions in China (WH, SH and HZ) were selected to show each 1D-G map.

In Fig. 14, one 2D-(A,G) China, three regions in China (WH, SH and HZ) and three regions in the USA (CA, NY and WA) were selected to show each 1D-A map. In Fig. 15, one 2D-(A,G) China, three regions in China (WH,

SH and HZ) and three regions in the USA(CA, NY and WA) were selected to show each 1D-G map.

In Fig. 16, four countries and their regions were put together in one page for convenience in direct visual comparisons.

1337 Genomes on C1 and C4 Genomic Index

Both C1 and C4 were used to process 1337 from Fig. 17 to Fig. 20 for four and eight countries, respectively.

In Fig. 17 on C1, one 2D map of all countries and five countries, the USA, China, Australia, Italy and France, were selected to show each 1D map. In Fig. 18 on C1, 2D and 1D maps of all countries and eight countries, the USA, China, Australia, Italy, Belgium, Brazil, Canada and Japan, were selected to show each 1D map.

In Fig. 19 on C4, eight countries and two countries, the USA and China were selected to show each 1D projection of various countries and regions for topolog- ical entropies, respectively. In Fig. 20 on C4, three countries, Australia, Italy and France, were selected to show each 1D projection of various regions for topological entropies.

Conclusion

Based on the corresponding relationship between phylogenetic trees and genomic index maps, four datasets of 20, 128, 381 and 1337 genomes for SARS-CoV-2 worldwide were used {C1, C2, C3, C4} .

It is interesting to see various clustering properties visualized on various genomic index maps to support further explorations for COVID-19 patients.

Genomic index schemes provide unique global invariants for genomes in general. Future activities will help people understand deeper mysteries in RNA viruses under complicated real-world environments.

Declarations

Conflict Interest: No conflict of interest has been claimed.

Acknowledgements The authors would like to thank NCBI,GISAID, CNGBdb, Nextstrain and Dr. Zhigang Zhang for providing invaluable information on the newest dataset collections of SARS-CoV-2 and other coronavirus genomes to support this project working smoothly.

References

1. HP Yao, XY Lu, ···, LJ Li, Patient-drived mutations impact pathogenicity of SARS-CoV-2, DOI: https://doi.org/10.1101/2020.04.14.20060160 https://www.medrxiv.org/10.1101/2020.04.14.20060160v2

2. Li C, Yang Y, Ren L. Genetic evolution analysis of 2019 novel coronavirus and coronavirus from other species. Infect Genet Evol. 2020 Mar 10;82:104285. doi: 10.1016/j.meegid.2020.104285. [Epub ahead of print] PMID: 32169673

3. Z. J. Zheng,Conjugate transformation of regular plan lattices for binary images, PhD Thesis, Monash University, 1994.

4. Jeffrey Z. J. Zheng, Christian H. H. Zheng, A framework to express variant and invariant functional spaces for binary logic, Frontiers of Electrical and Electronic Engineering in China, 5(2):163-172, Higher Educational Press and Springer-Verlag, 2010.

5. Jeffrey Z.J. Zheng, Christian H.H. Zheng and Tosiyasu L. Kunii. A Framework of Variant Logic Construction for Cellular Automata, Cellular Automata - Innovative Modeling for Science and Engineering, Dr. Alejandro Salcido (Ed.), InTech Press, 2011.

6. Jeffrey Zheng, Variant Construction from Theoretical Foundation to Applications, Springer Nature 2019 https://www.springer.com/in/book/9789811322815

7. Jeffrey Zheng, Variant Construction Theory and Applications, Vol. 1: Theoretical Foundation and Applications, Science Press 2020 (Chinese, Formal Publishing Soon). 郑智捷, 变值体系 理论及其应用 第 1 册: 理论基础及其应用，科学出版社 20 (即将正式发行)

8. Jeffrey Zheng, ResearchGate: http://researchgate.net/pprofile/Jeffrey Zheng

9. Jeffrey Zheng, Chris Zheng, Biometrics and Knowledge Management Information Systems, Chapter 11: Variant Construction from Theoretical Foundation to Applications, Springer Nature 2019, 193-202 https://link.springer.com/chapter/10.1007/978-981-13-2282-2 11 被斯普林格–自然杂志出版社，选入抗击新型冠状病毒肺炎研究(Research of COVID-19)资料汇集。推荐给 PMC 和 WHO（PubMed Central PMC and the World Health Organization WHO）以方便全球科学研究人员免费使用

Figures

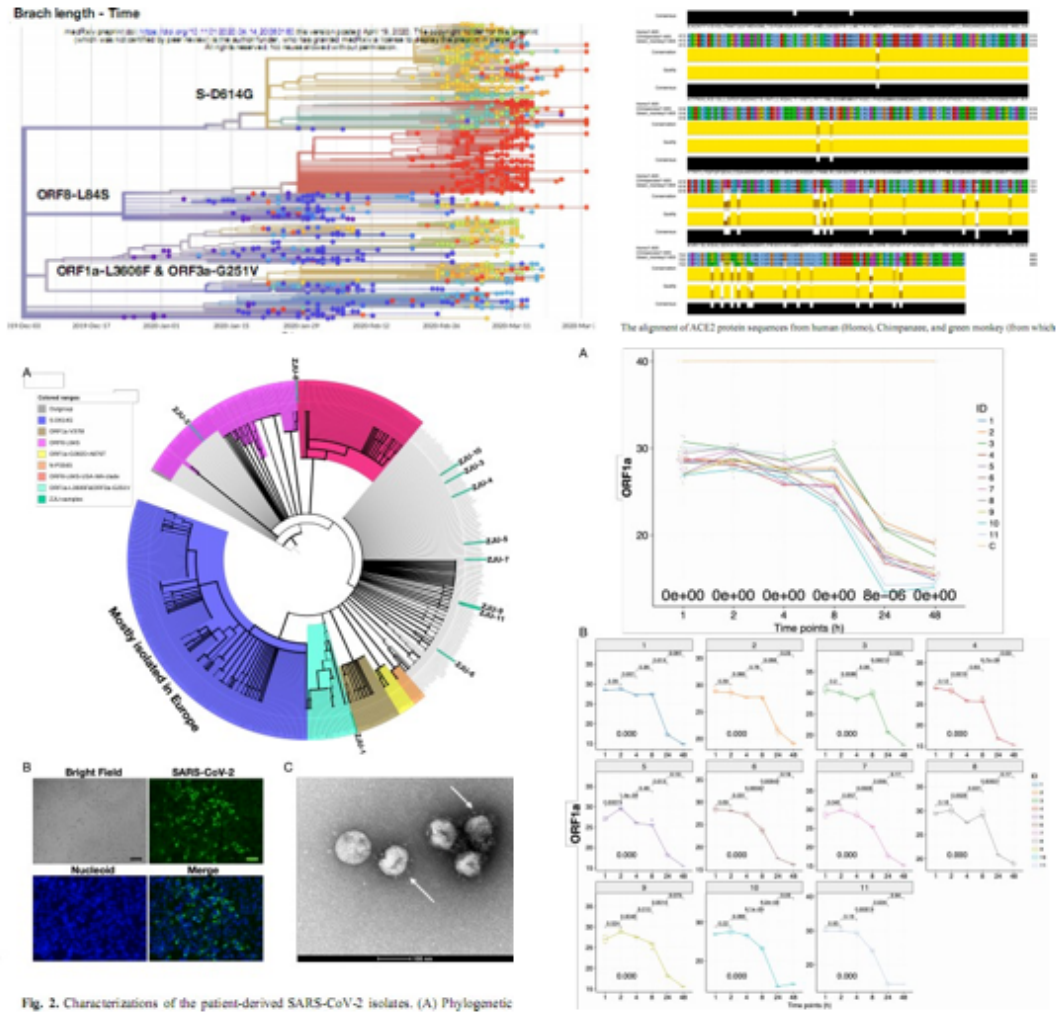


Fig. 2. Characterizations of the patient-derived SARS-CoV-2 isolates. (A) Phylogenetic

Figure 1

Key steps of phylogeny used in viral toxicologic tests for COVIP-19 patients from [1]

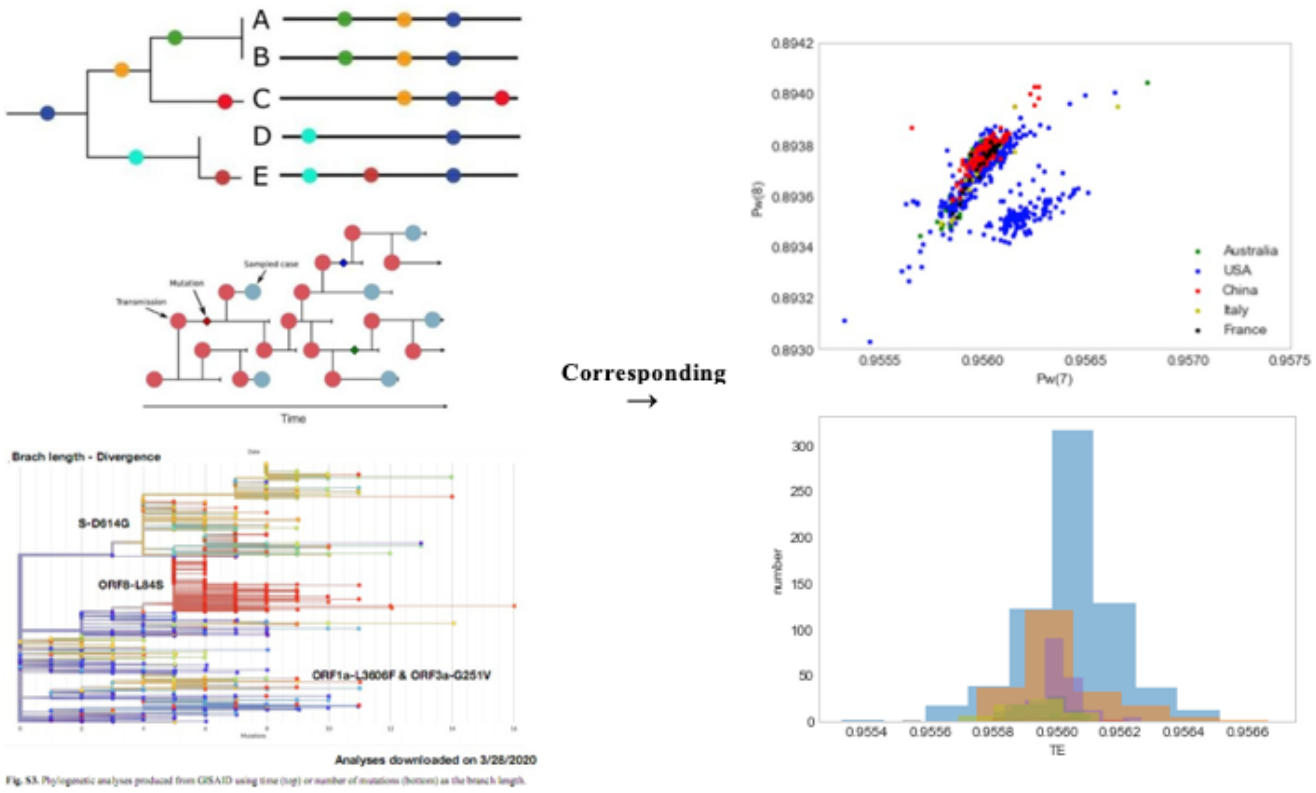


Figure 2

Phylogeny tree and density distributions of various genomic index maps

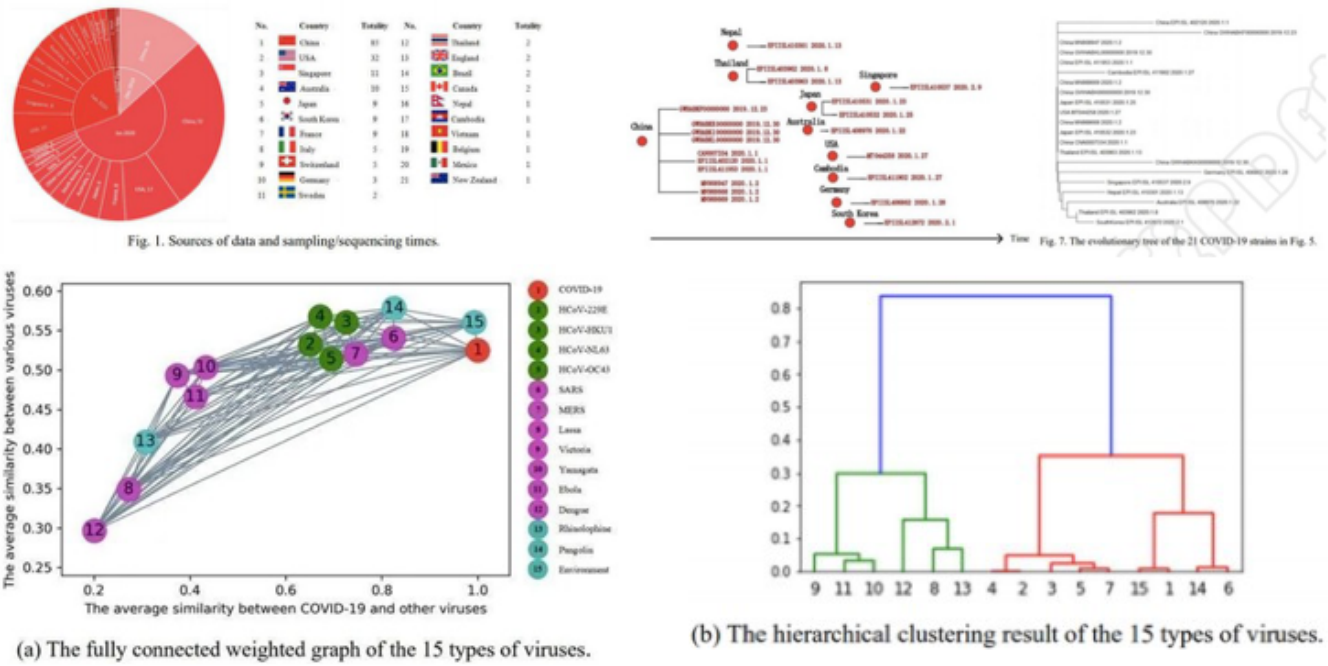


Figure 3

Main steps of multiple coronaviruses on similarity networks [2]

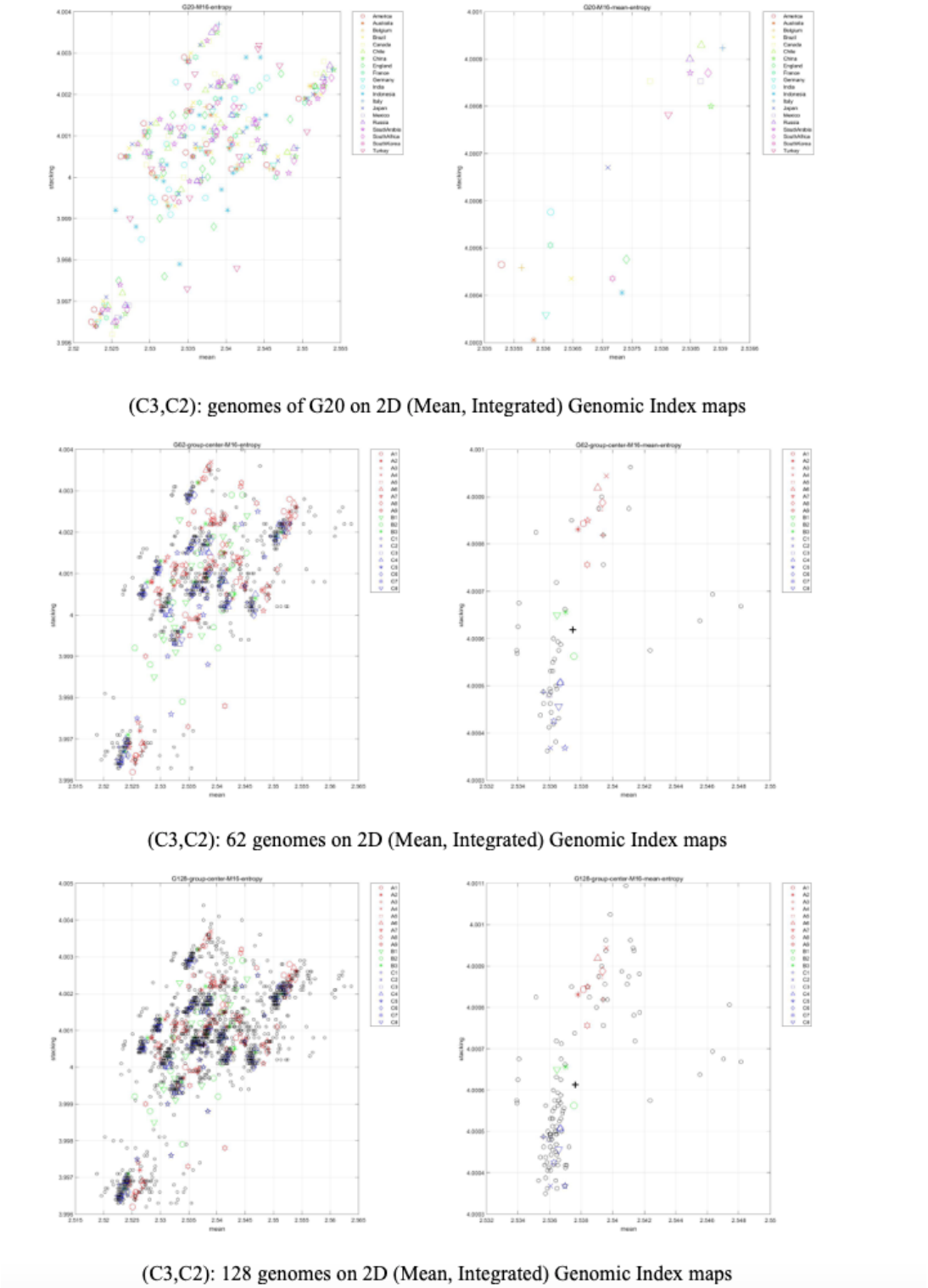
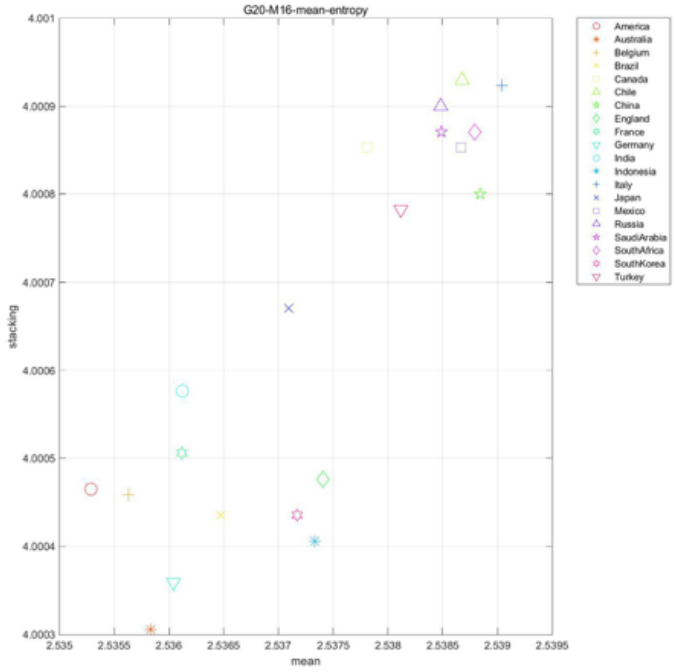
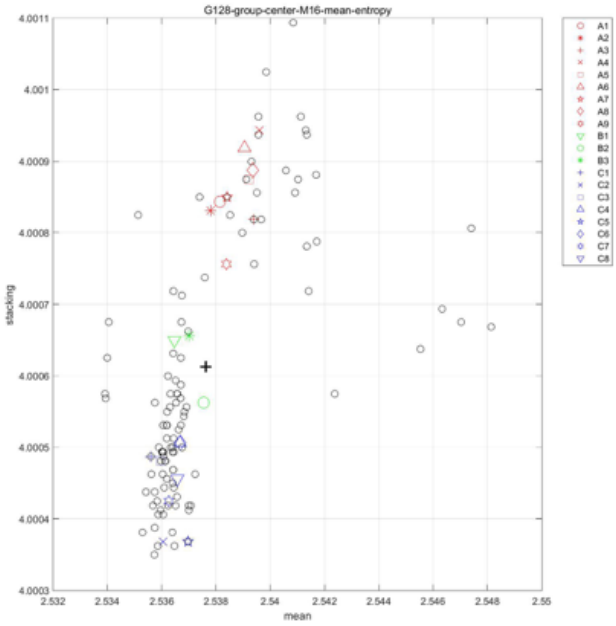


Figure 4

(C3,C2): 128 genomes on 2D (Mean, Integrated) Genomic Index maps



(a) G20 Regions



(b) 128 Samples

Figure 5

(C3,C2): Genomes from G20 to 128 Samples on 2D (Mean, Integrated) Genomic Index Maps; (a) G20; (b) 128 Genomes

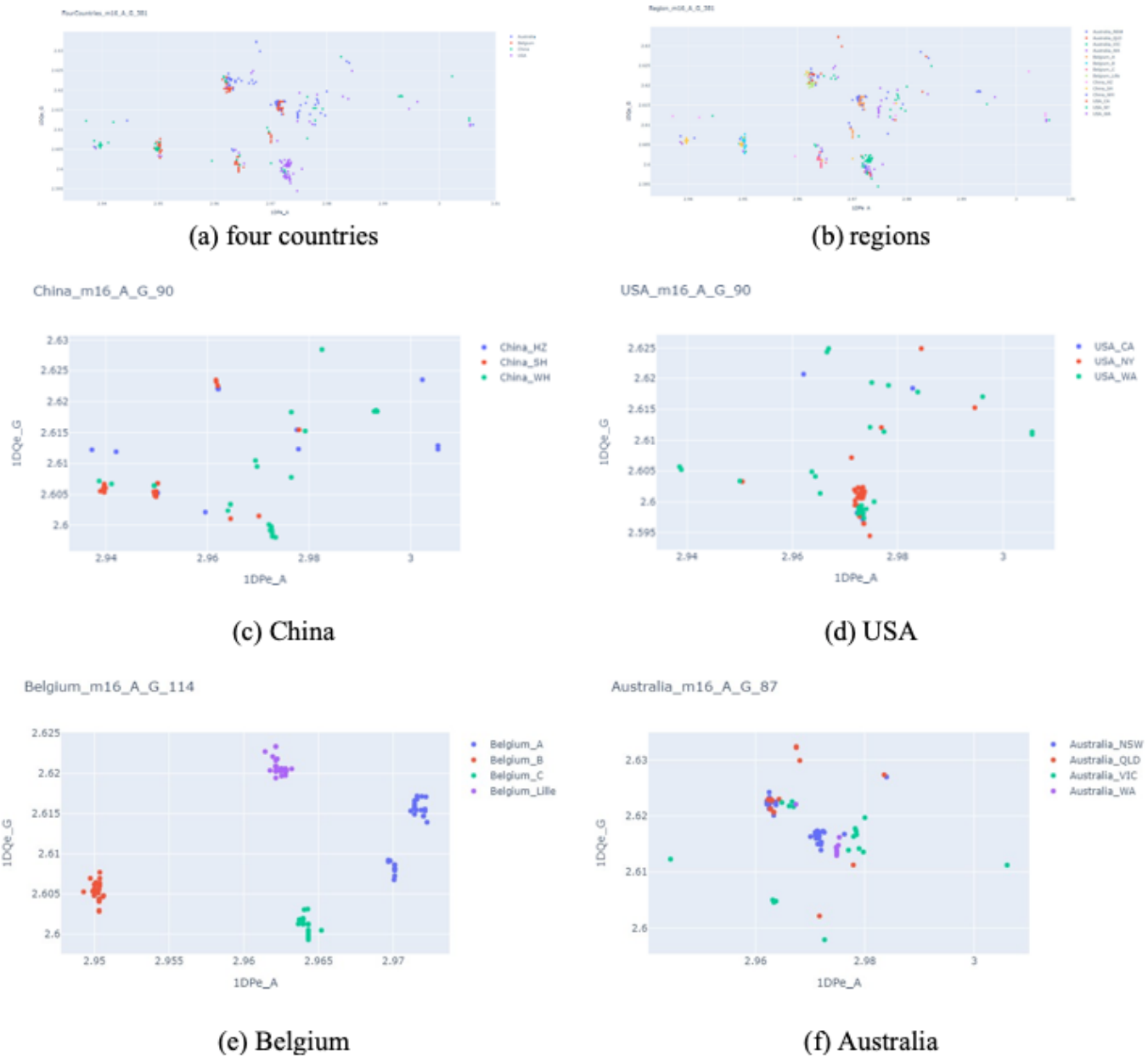


Figure 6

C1: 2D-(A,G) maps, four countries, Australia, Belgium, China, and USA; (a) 2D four coun- tries, (b) four countries of regions, (c) China, (d) USA, (e) Belgium, (f) Australia

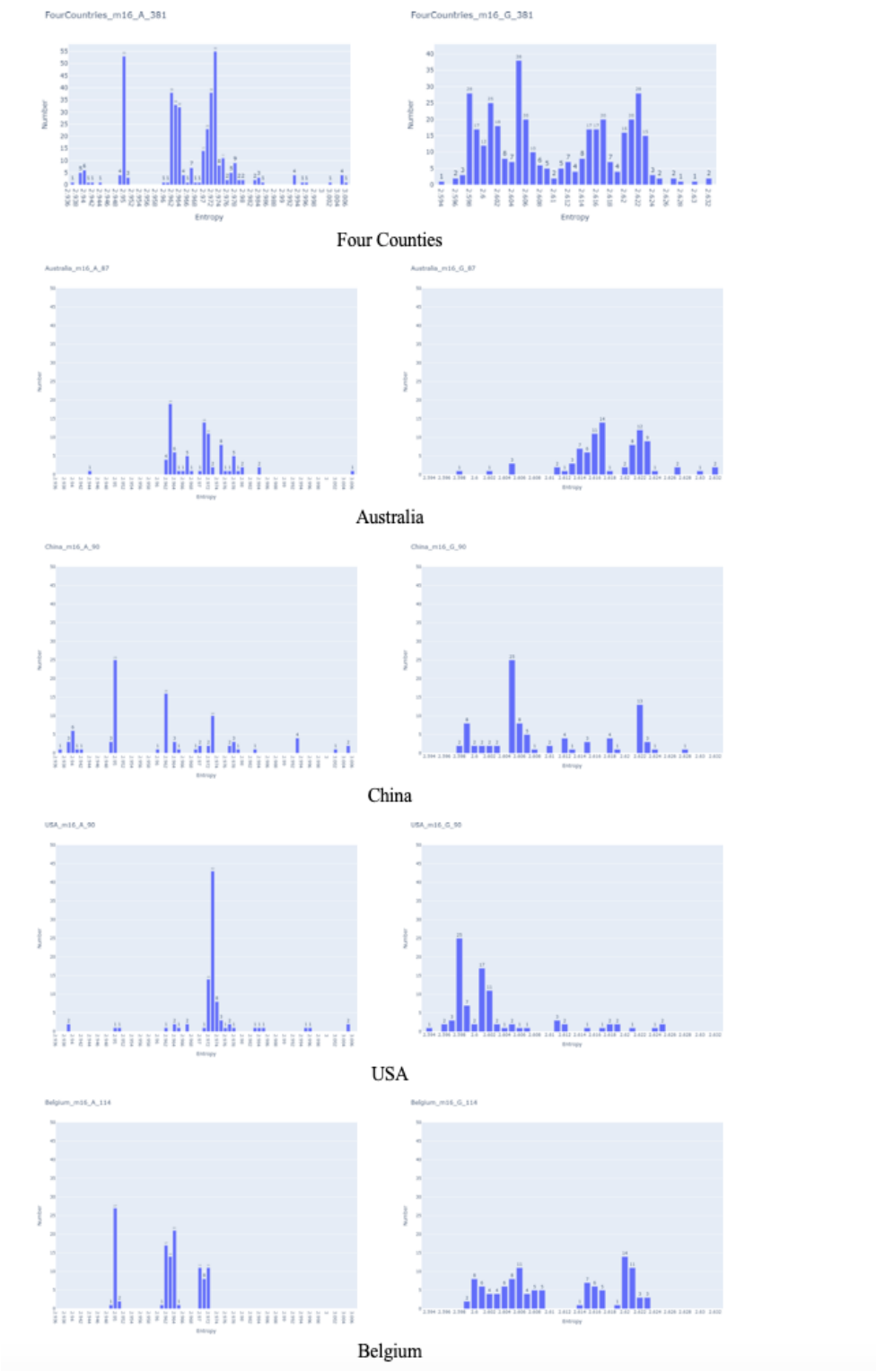
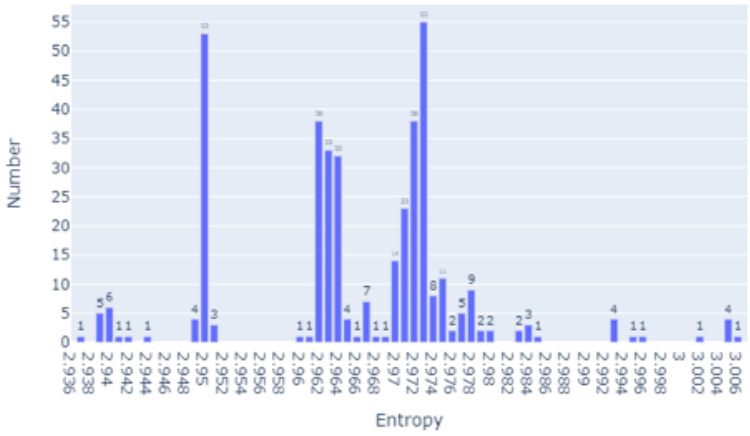


Figure 7

C1: 381 SARS-CoV-2 genomes of four countries in 1D projections of combinatorial ge- nomic index maps; left 1D-A; right 1D-G

FourCountries_m16_A_381



FourCountries_m16_G_381

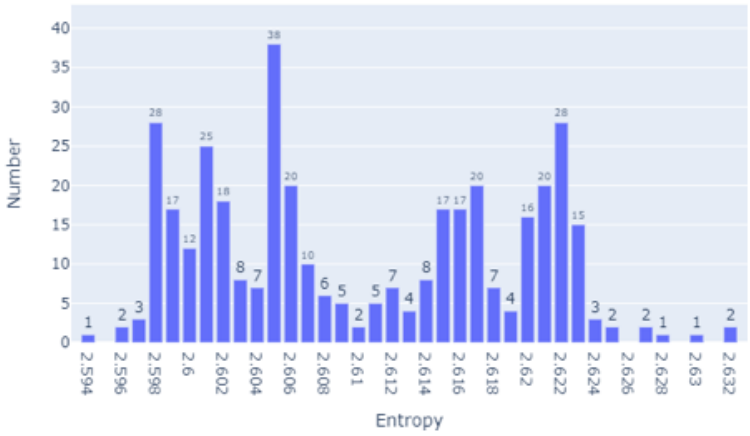
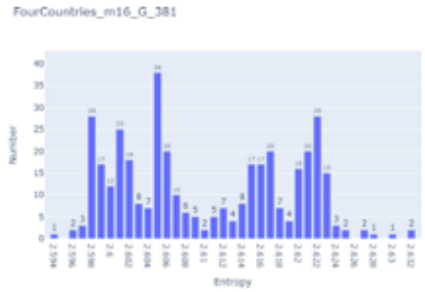


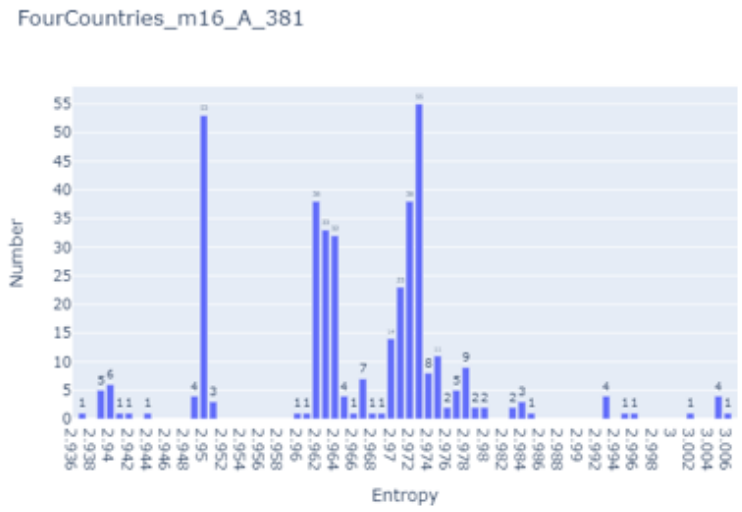
Figure 8

C1: 381 SARS-CoV-2 genomes on 2D combinatorial genomic index maps and two 1D projections

(a) 2D-(A, G) map, four countries



(c) 1D-G Y-Projection



(b) 1D-A X-Projection

Figure 9

C1: 2D-(A,G) and 1D maps, four countries, Australia, Belgium, China, and USA; (a) 2D- (A,G); (b) 1D-A X axis; (c) 1D-G Y axis

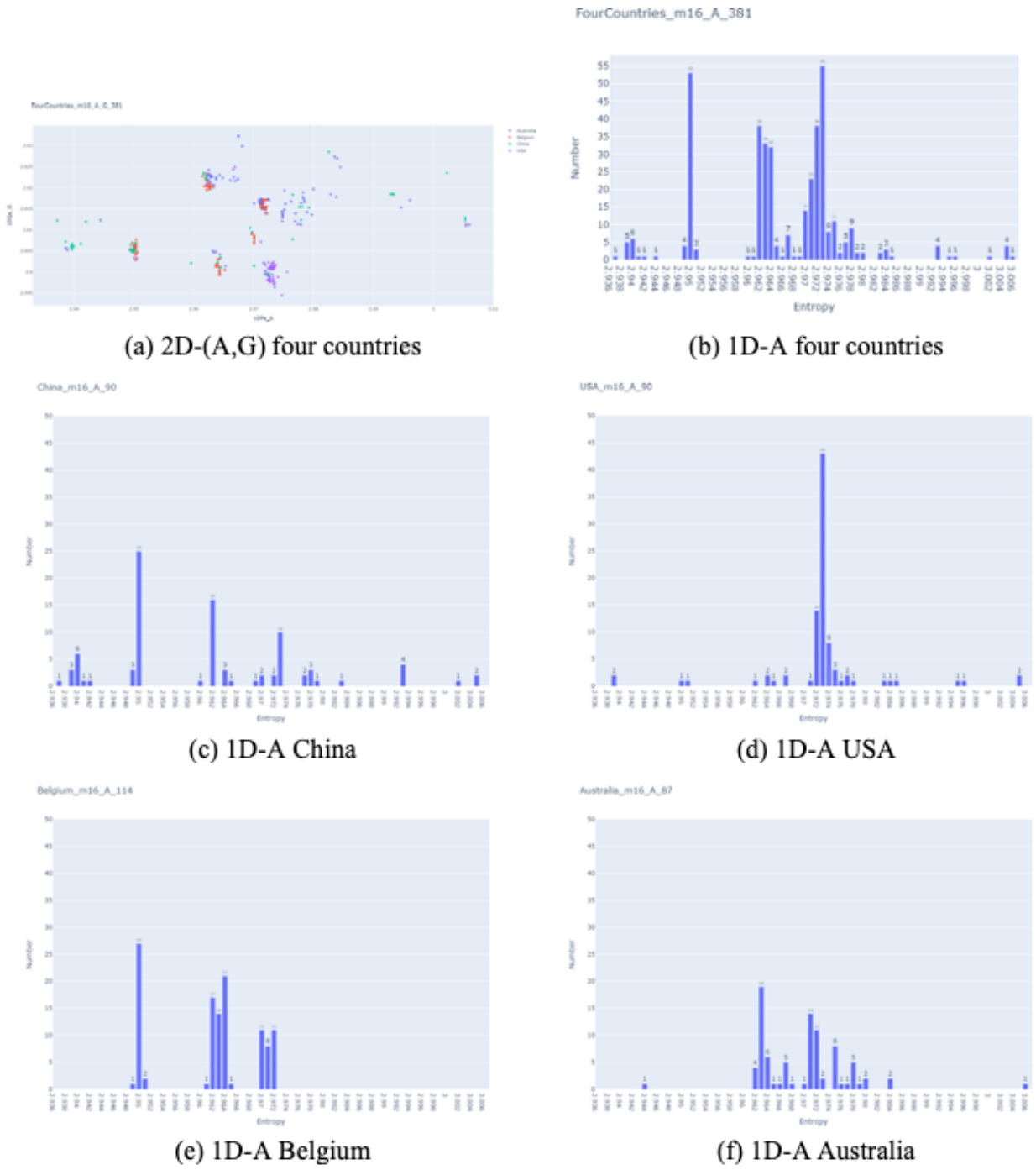
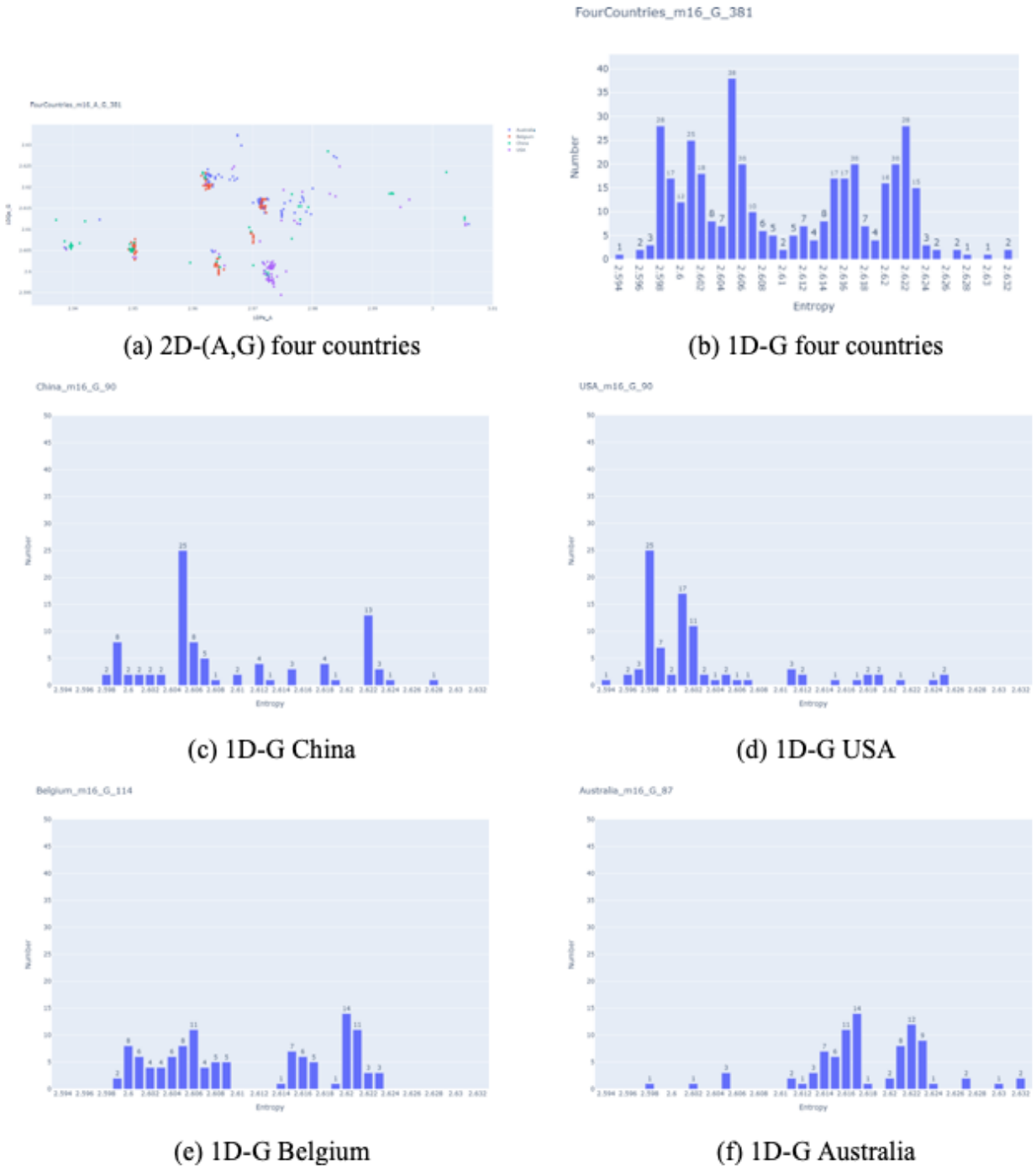


Figure 10

C1: 2D-(A,G) & 1D-A maps, four countries, Australia, Belgium, China, and USA; (a) 2D four countries, (b) 1D-A four countries, (c) China, (d) USA, (e) Belgium, (f) Australia



(a) 2D-(A,G) four countries

(b) 1D-G four countries

(c) 1D-G China

(d) 1D-G USA

(e) 1D-G Belgium

(f) 1D-G Australia

Figure 11

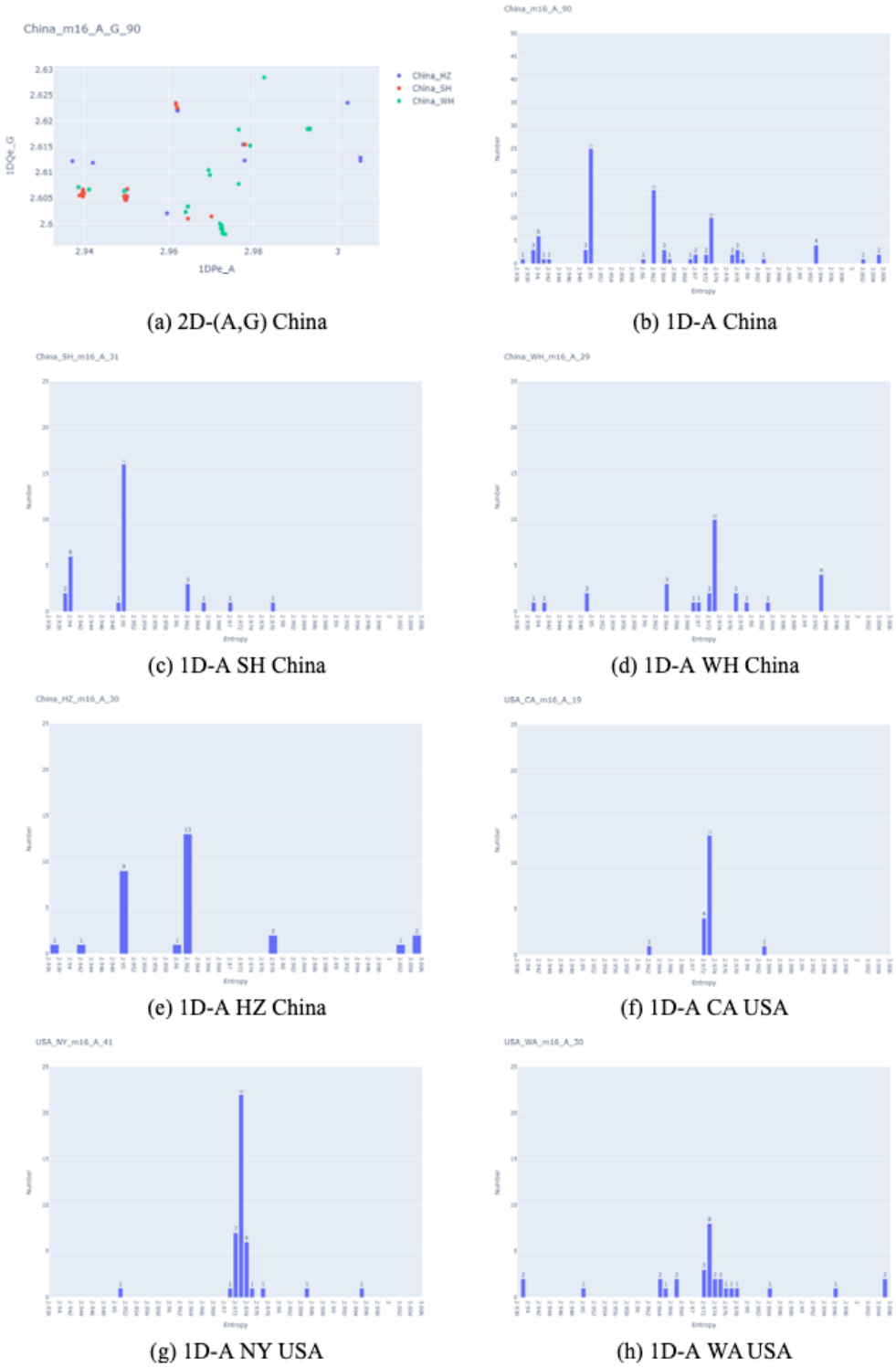
C1:2D-(A,G) and 1D-G maps, four countries, Australia, Belgium, China, and USA; (a) 2D four countries, (b) 1D-A four countries, (c) China, (d) USA, (e) Belgium, (f) Australia

Figure 12

C1: 2D-(A,G) & 1D-A maps, three regions in the USA (CA, NY, WA); Three Regions in China (WH, SH, HZ); (a) 2D-(A,G), USA, (b) 1D-A USA, (c) CA USA, (d) NY USA, (e) WA USA (f) WH China (g) SH China (h) HZ China

Figure 13

C1: 2D-(A,G) and 1D-G maps, three regions in the USA (CA, NY, WA); Three Regions in China (WH, SH, HZ); (a) 2D-(A,G) USA, (b) 1D-A USA, (c) CA USA, (d) NY USA, (e) WA USA (f) WH China (g) SH China (h) HZ China



(a) 2D-(A,G) China (b) 1D-A China

(c) 1D-A SH China (d) 1D-A WH China

(e) 1D-A HZ China (f) 1D-A CA USA

(g) 1D-A NY USA (h) 1D-A WA USA

Figure 14

C1: 1D-A maps, three regions in China (HZ, SH, WH); three regions in the USA (CA, NY, WA); (a) 2D China, (b) 1D-A HZ China, (c) SH China, (d) WH China, (e) HZ China, (f) CA USA, (g) NY USA (h) WA USA

(a) 2D-(A,G) China

(b) 1D-G China

(c) 1D-G SH China

(d) 1D-G WH China

(e) 1D-G HZ China

(f) 1D-G CA USA

(g) 1D-G NY USA

(h) 1D-G WA USA

Figure 15

C1: 1D-G Maps, Three Regions in China: HZ, SH, WH; Three Regions in USA: NY, WA; (a) 2D China, (b) HZ China, (c) SH China, (d) WH China, (e) HZ China, (f) CA USA, (g) NY USA (h) WA USA

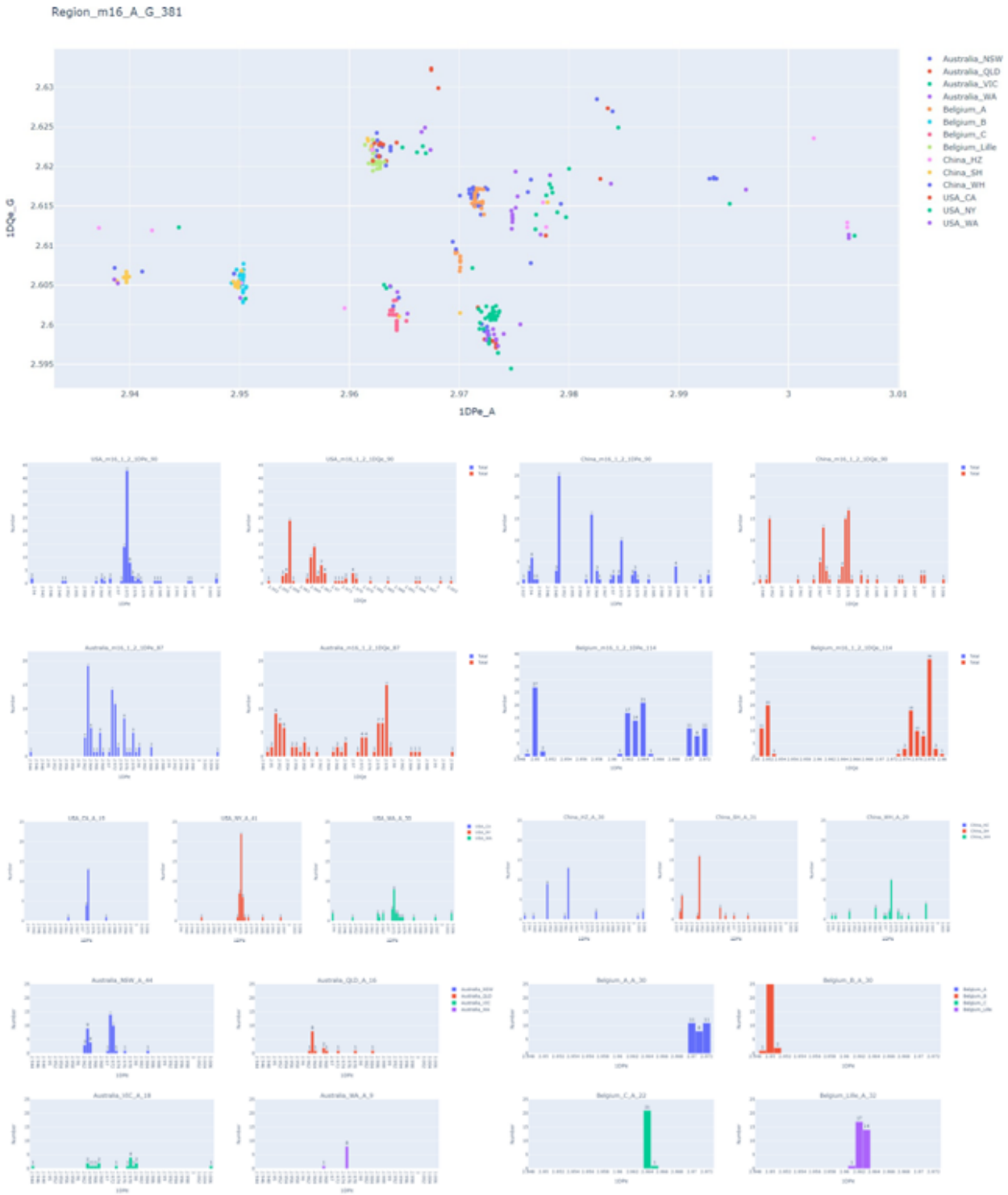


Figure 16

C1: 381 SARS-CoV-2 genomes of four countries on various regions in 2D and 1D projections of combinatorial genomic index maps

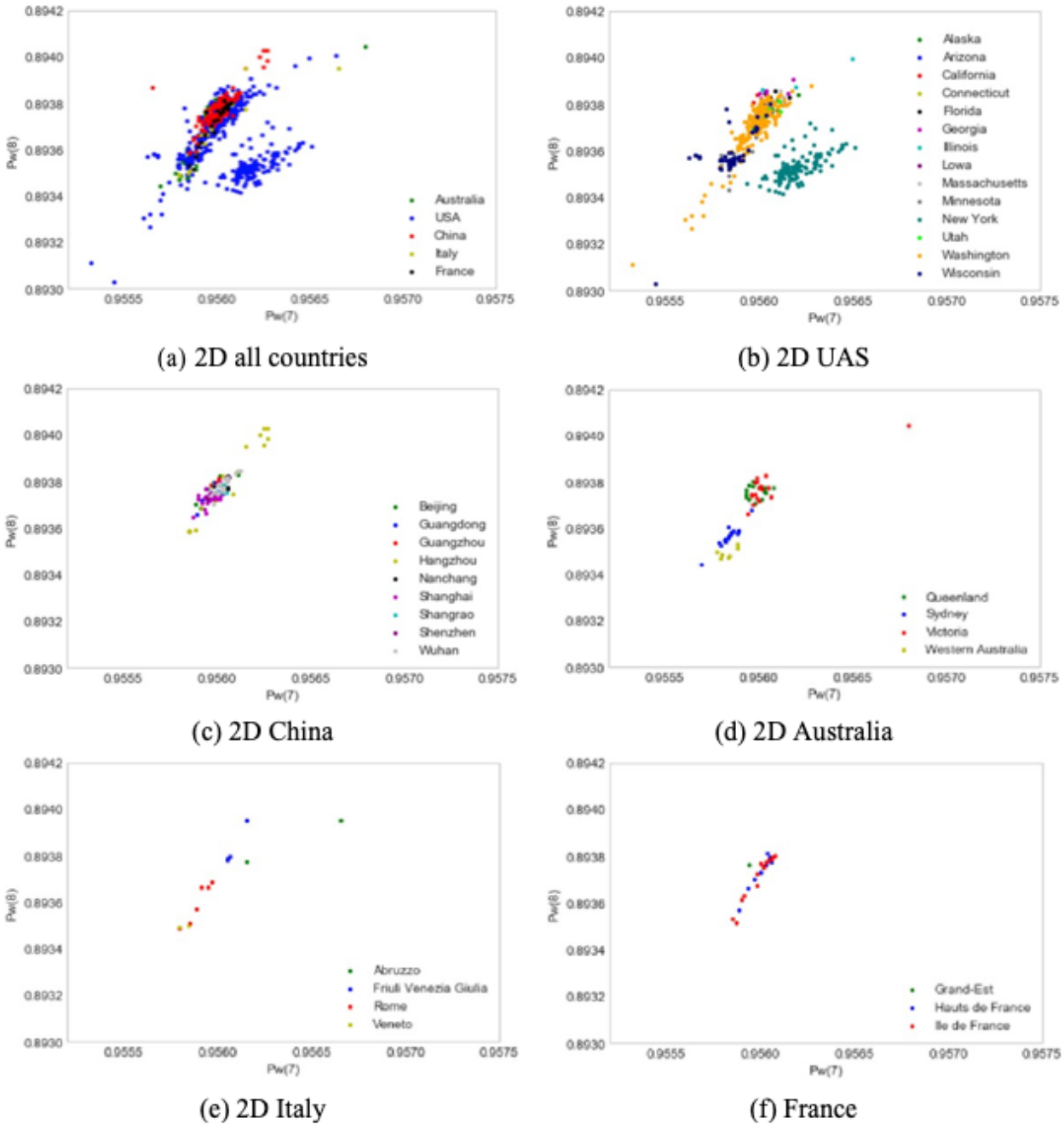


Figure 17

C4: 1337 genomes on 2D genomic index maps in topological entropies

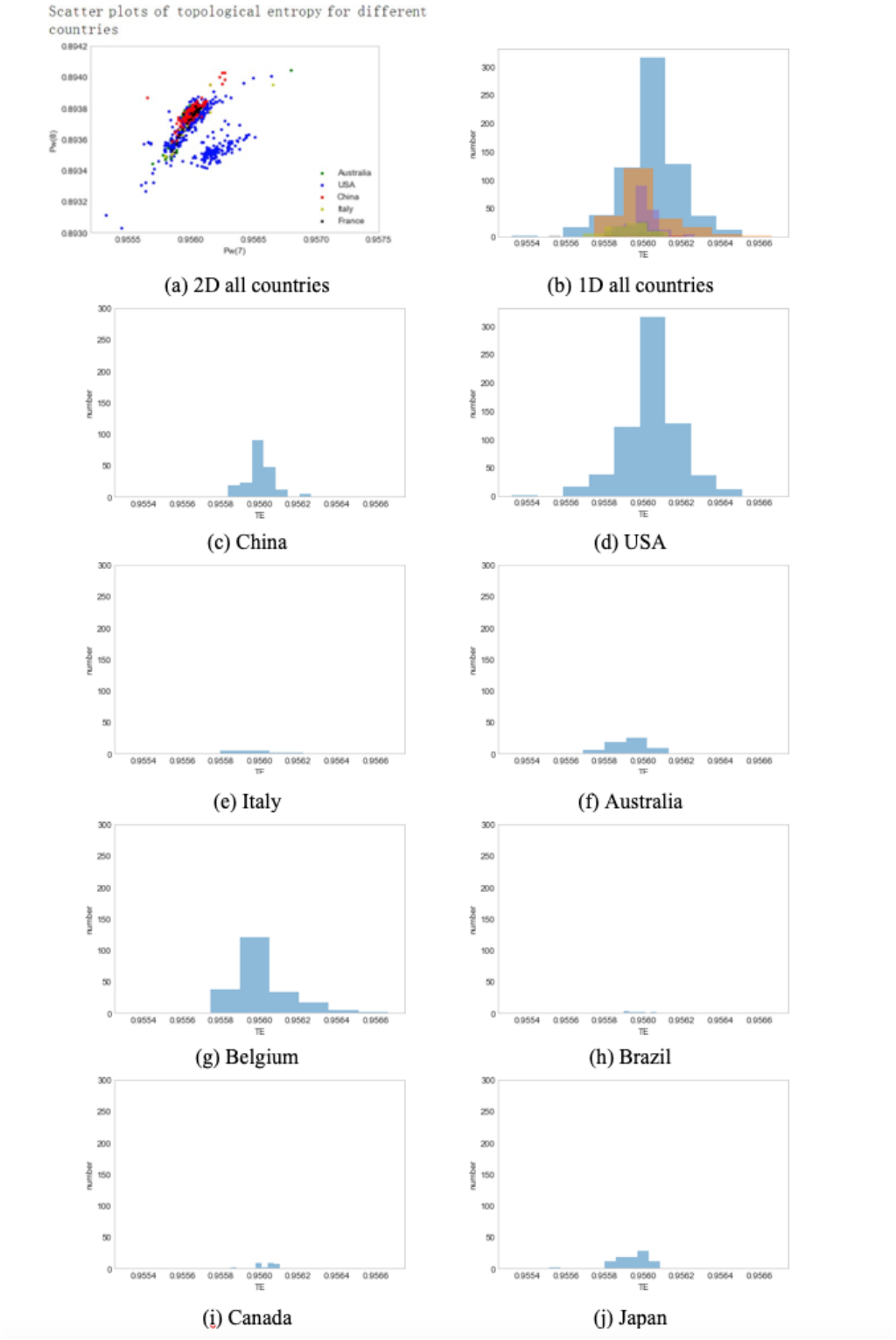


Figure 18

C4: 1337 genomes on topological entropies in 2D and 1D genomic index maps

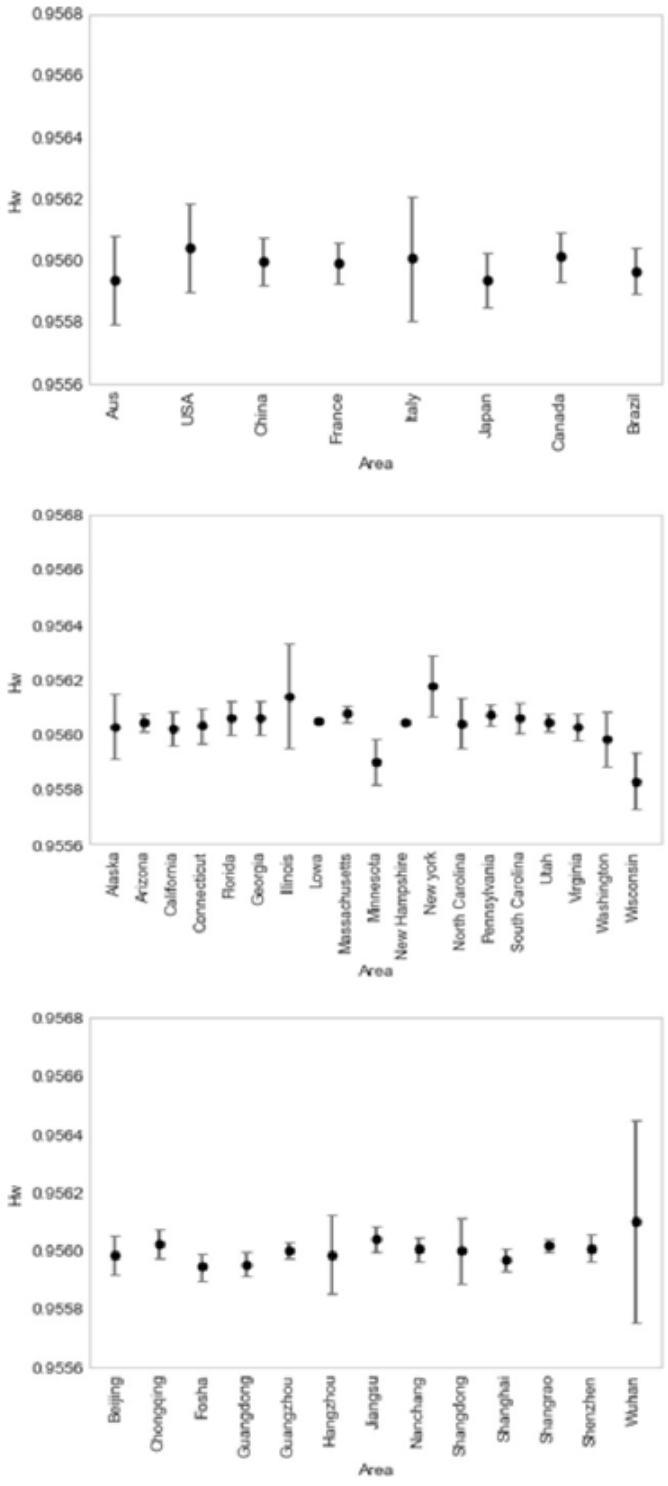


Figure 19

C4: 1337 genomes on topological entropies in the USA and China on 1D projections of genomic index maps

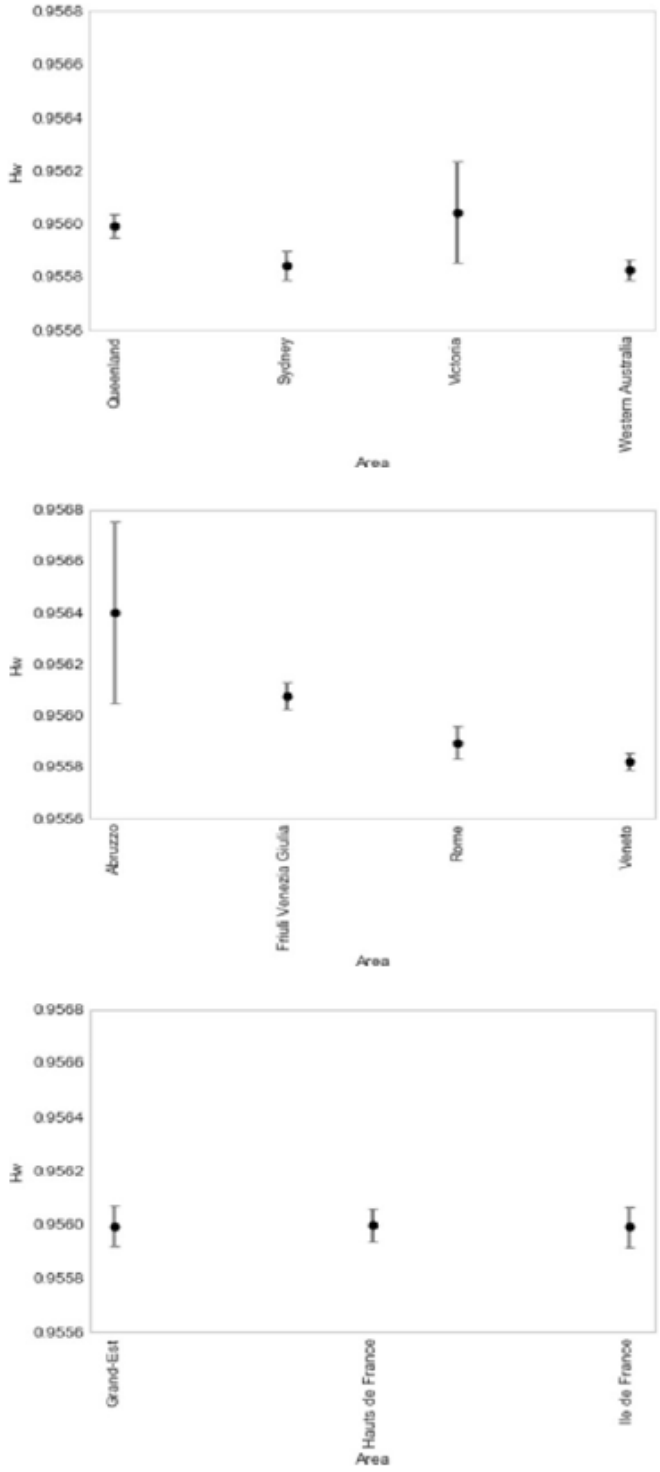


Figure 20

C4: 1337 genomes on topological entropies in Australia, Italy, and France on 1D projec- tions of genomic index maps